LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

UCRL-JRNL-205731

Evolution and Functional Classification of Vertebrate Gene Deserts

I. Ovcharenko, G.G. Loots, M.A. Nobrega, R.C. Hardison, W. Miller, L.J. Stubbs

August 5, 2004

Genome Research

Evolution and Functional Classification of Vertebrate Gene Deserts

UCRL: UCRL-JRNL-205731

IM: #309641

Evolution and Functional Classification of Vertebrate Gene Deserts

Ivan Ovcharenko[1,*], Gabriela G. Loots[2], Marcelo A. Nobrega[3], Ross Hardison[4], Webb Miller[5,6] and Lisa Stubbs[2]

[1]Energy, Environment, Biology and Institutional Computing, Lawrence Livermore National Laboratory, Livermore, CA 94550

[2]Genome Biology Division, Lawrence Livermore National Laboratory, Livermore, CA 94550

[3]Department of Genome Sciences, Lawrence Berkeley National Laboratory, Berkeley, CA 94720

[4]Department of Biochemistry and Molecular Biology, The Pennsylvania State University, University Park, PA 16802

[5]Department of Computer Science and Engineering, The Pennsylvania State University, University Park, PA 16802

[6]Department of Biology, The Pennsylvania State University, University Park, PA 16802

*corresponding author

Phone: (925) 422-5035

Fax: (925) 422-2099

Email: ovcharenko1@llnl.gov

ABSTRACT

Gene deserts, long stretches of DNA sequence devoid of protein coding genes, span approximately one quarter of the human genome. Through human-chicken genome comparisons we were able to characterized one third of human gene deserts as evolutionarily stable - they are highly conserved in vertebrates, resist chromosomal rearrangements, and contain multiple conserved noncoding elements physically linked to their neighboring genes. A linear relationship was observed between human and chicken orthologous stable gene deserts, where the human deserts appear to have expanded homogeneously by a uniform accumulation of repetitive elements. Stable gene deserts are associated with key vertebrate genes that construct the framework of vertebrate development; many of which encode transcription factors. We show that the regulatory machinery governing genes associated with stable gene deserts operates differently from other regions in the human genome and relies heavily on distant regulatory elements. The regulation guided by these elements is independent of the distance between the gene and its distant regulatory element, or the distance between two distant regulatory cassettes. The location of gene deserts and their associated genes in the genome is independent of chromosomal length or content presenting these regions as well-bounded regions evolving separately from the rest of the genome.

INTRODUCTION

The sequence of the human genome provides researchers with the substrate upon which the code of life is embedded. One of the main challenges of the post-genome sequencing era is to understand how the genetic code is organized, and especially, what kinds of factors contribute to its complex and precise coordination of gene expression. Multiple mechanisms have already been recognized to influence genomic organization and function either in the form of *cis*-regulatory sequences controlling gene expression, or barrier elements defining physical domains that anchor genomic regions to specific nuclear localizations (ref). Nevertheless, attention is now shifting to the impact that other aspects of genomic architecture may have on genome function. To this end, it is known that vertebrate genomes are highly heterogeneous, with the human genome, for example, being surprisingly rich in 'junk' DNA inhabited by repetitive elements that together account for about half of the genomic sequence (Lander et al. 2001).

One of the most baffling genomic architectural asymmetries described upon the sequencing of the human genome concerns the uneven distribution of genes throughout the genome. It is estimated that ~25% of the human genome is represented by *gene deserts* – long regions that do not code for proteins and do not have any obviously defined functions. Recently, a coupled computational and experimental search for functional activities within *gene deserts* identified several distant gene regulatory elements embedded in a 800kb noncoding interval flanking the *DACH* gene (Nobrega et al. 2003). Despite pinpointing to key transcriptional regulatory elements present within the *DACH gene desert*, these observations didn't address fundamental issues about what purpose do *gene deserts* serve in general, and why do they persist in vertebrate genomes? Furthermore, contrary to results obtained in the *DACH* desert study recent observations

have suggested that some *gene deserts* are potentially nonessential to genome function (Pennisi, Science). It is possible that these inconsistencies in fact reflect the existence of distinct categories of *gene deserts*, with some deserts harboring sequence elements with critically important and conserved biological roles and others not. In order to investigate this possibility, we have applied a comparative sequence analysis strategy to identify *gene deserts* across the vertebrate subphylum in order to describe and evolutionarily characterize these peculiar genomic intervals. Specifically, we focused on sequence comparisons with the chicken genome, a valuable organism strategically positioned between rodents and fish in the vertebrate evolutionarily tree. By analyzing the genomic structure, the conservation patterns, and the evolutionary relationships of gene deserts we were able to classify gene deserts into two functionally different groups and to provide some insights regarding the functions of these intervals in the human genome.

RESULTS

Human Gene Deserts

In order to identify *gene deserts* we scanned the human genome (NCBI Build 34; UCSC freeze hg16) and identified 18,134 intergenic regions as defined by the *knownGene* gene annotation (Karolchik et al. 2003). These genomic intervals span 61.2% of the human genome, where telomeric and centromeric regions (defined as unsequenced intervals longer than 250kb) were excluded from the analysis. Intergenic regions were separated into two different categories – regular intergenic (intergenic regions ranging in size from 25% to 50% percentiles of the length distribution curve) and *gene deserts* (top 3% longest intergenic intervals). Regular intergenic regions ranged from 5.6kb to 21.5kb. A total of 545 intergenic regions were classified as *gene deserts,* varying in length between 638kb and 5.1Mb. We found that ~25% of euchromatin is enclosed within *gene deserts,* consistent with previous estimates (Nobrega et al. 2003; Venter et al. 2001). Two small human chromosomes (HSA17; HSA19) are distinct outliers carrying a disproportional high amount of regular intergenic regions, and with few exceptions, lacking *gene deserts*. In contrast, HSA13, 4 and 5 are heavily populated with *gene deserts*, which cover up to 40% of the length of each chromosome (Figure 1). These observations correlate with the gene-rich makeup of human chromosomes 17 and 19, and the relatively gene-sparse nature of the chromosomes 4, 5 and 13 (Dunham et al. 2004; Grimwood et al. 2004)

To identify putative signatures that define gene deserts, we carried out a comprehensive comparison between regular intergenic intervals and *gene deserts*, contrasting characteristics of these genomic segments to those of the entire human genome and to gene-rich regions. In order to define gene-rich regions we first identified

all the gene clusters in the human genome separated by intergenic regions longer than 100kb. Out of the 3,581 clusters fitting these criteria, 144 clusters contained 20 or more genes. The three most gene-rich regions were located on HSA19, HSA17 and HSA16, each spanning over 4Mb of sequence and comprising more than140 genes. These gene-rich regions have partially originated through the expansion of zinc-finger transcription factors, Kallikreins, Keratins, and other tandem duplications of gene families (Dehal et al. 2001; Shannon et al. 2003), but these regions are also densely packed with unique genes of many different types. In total, these *gene-rich* regions covered 285Mb of the human genome with 15 clusters originating from the most gene-rich human chromosome HSA19.

To address the functional significance of *gene deserts* in the human genome we quantified several parameters that might reveal signatures unique to *gene deserts* (Table 1). In contrast to other genomic regions, *gene deserts* revealed a significantly elevated density of single nucleotide polymorphisms (SNPs), a decrease in sequence similarity to chicken and mice, and a low GC content. These results suggest reduced purifying selection pressure may be operating in general on *gene desert* regions, furthering the hypothesis that *gene deserts* may represent primarily biological wastelands full of pseudogenes, repeats and other similar nonfunctional sequences. Contrary to this hypothesis, the fraction of *gene deserts* corresponding to repetitive sequences is not higher (50.5%) than that found in regular intergenic intervals (51.9%). Moreover, it has been shown that some human *gene deserts* harbor distant gene regulatory elements that are deeply conserved down to fish (Nobrega et al. 2003).

The puzzle posed by the paradox that *gene deserts* in the human genome can serve both as junkyards and oasis of functional noncoding elements can be tackled through

sequence comparisons with the chicken and mouse genomes. Despite the fact that different categories of genomic regions display similar average values of repeat content and conservation parameters in general, we observed a wide distribution of values within each one of these categories. For example, there are *gene deserts* that are as repeat-rich as 90% or repeat-poor as 30%; some *gene deserts* are completely diverged from their mouse and chicken counterparts while others are highly conserved, with ~43% (human/mouse) and ~12% (human/chicken) of non-repetitive sequence blocks corresponding to evolutionarily conserved regions (ECRs). Assuming that the level of noncoding sequence conservation reflects the level of purifying selection, and the density of repetitive elements indicate neutrally evolving regions, there should be a negative correlation between repeat content and the level of evolutionary conservation of non-repetitive sequences. Nevertheless, this correlation can be clearly observed only for a minor subset of human *gene deserts* that are outliers in either human-chicken conservation or repeat content (Figure 2A).

Categories of Gene Deserts

Human *gene deserts* fall into two distinct categories depending on the degree of sequence conservation to the distantly related chicken genome: *stable gene deserts* (172 regions) have >2% of non-repetitive sequence conserved to chicken and *flexible gene deserts* (373 regions) that have <2% conserved. Most *stable gene deserts* span a very narrow window in repeat content distribution with the average of 47.0% density of repetitive elements. This value is lower then the average repeats density across the whole genome or in *gene-rich* regions suggesting an elevated purifying selection pressure applied to *stable gene deserts*. In sharp contrast to the human-chicken conservation level,

human-mouse comparisons do not differentiate between *stable* and *flexible gene deserts* (Figure 2B). Therefore, the human-mouse sequence similarity profile cannot reliably predict whether a *gene desert* will be conserved in chickens or not. It is also worth mentioning that the previously described *DACH gene desert*, which contains several experimentally defined transcriptional gene regulatory elements, is one of the most highly conserved *stable gene desert* with ~37% human-mouse conservation level and one of the lowest repeat content.

We observed that a large number of *stable gene deserts* appear contiguously in neighboring genomic segments, separated from each other by small gene clusters. By searching for *stable gene deserts* that are separated by <1Mb of sequence including 3 or fewer gene transcripts, we identified 56 *neighboring stable gene deserts* (33% out of 172). The identified dataset of genes interspersed between *stable gene deserts* allowed us to characterize a particular class of genes that have evolved in a purely noncoding genomic environment. Gene Ontology (GO) functional characterization of these genes indicated enrichment in genes involved in transcriptional gene regulatory functions and a depletion in genes implicated in the "response to stimulus" category (p-value < 1e-3 as obtained through the comparison with the purely-by-chance expectation). The latter indicates that gene with species specific functions do not need to be preserved through the evolution of vertebrates, but instead would benefit from the evolutionarily changes. This gene bias towards transcription factors indicate that not only the transcriptional machinery is highly preserved through the evolution of vertebrates, but the regulation code for these transcription factors is kept under high purifying selection as well. Some other categories observed in this gene dataset included: skeletal development (BMP2), electron transport (COX7A3), muscle development (MEF2C), calcium ion binding

(DGKB), apoptosis (FKSG2), and cell cycle (DBC1). The majority of these groups represent genes involved in the most critical developmental steps and essential biochemical processes of vertebrates.

We also investigated the GO characterization for all the genes flanking *gene deserts*. While the enrichment was not very striking when all *gene deserts* were analyzed, it highlighted some well-defined categories for the *stable gene deserts* subcategory. Namely, we observed enrichment in genes coding for transcription factors, genes involved in the regulation of transcription and DNA binding, genes participating in regulation of metabolism and development (Table 2). Drastically different functional specification was observed for *flexible gene deserts* pointing to genes with lineage-specific functions. Genes involved in intercellular communication processes, receptor activity, neurophysiological processes and organogenesis were found to be enriched in regions flanking the *flexible gene deserts* (Table 2). Similar to genes flanked on both sides by *stable gene deserts*, this analysis predicts that the transcripts associated with *stable gene deserts* partake in core biological processes of vertebrate organisms. This observation compounds to previously reported observations that in invertebrates genes involved in organism development or encoding for transcription factors are surrounded by much larger intergenic sequences than housekeeping genes or genes involved in metabolism (Nelson et al. 2004). These data also suggest that these genes, often endowed with complex expression patterns are likely regulated by multiple regulatory units, which during vertebrate evolution and genome size expansion have drifted apart from their target gene, creating some of the stable gene deserts now present in vertebrate genomes.

Primate Evolution Of The Gene Deserts

Although the relative density of repetitive elements in *gene deserts* is comparable with the average distribution in the genome and is slightly lower than the average intergenic interval, the content of the various classes of repetitive elements varies in *gene deserts*. The density of LINE elements is significantly higher while the density of SINE elements is much lower in *gene deserts*, when compared to averages for the entire human genome (Figure 3). This characteristic trend is present for both *stable* and *flexible gene deserts*. An opposite and more profound effect is observed for regular intergenic and *gene-rich* regions – the level of SINE elements is increased and the LINE element content is decreased. Ancient L2 repetitive elements contribute only minimally to the distribution of repeats (3.3% of the overall distribution in average) and do not have a pronounced dependence on different genomic categories – the backbone of Figure 3 distribution mainly consists of L1 repetitive elements. L1 elements represent a recent class of repeats that establish a solid presence in higher vertebrate genomes since the time following the amphibian and avian radiation, but preceding mammalian radiation. This suggests that the observed imbalance of LINE vs SINE repetitive elements populating *gene deserts* corresponds to the effects of the post-rodent evolution of the human genome, highlighting differences in recent evolutionary events of these two genomes.

These large differences in categories of repetitive sequences in various genomic fractions suggest a purifying selection against accumulation of SINE elements in *gene deserts* and LINE elements in regular intergenic intervals and *gene-rich* regions. A possible explanation for this selective pressure preventing SINE accumulation in *gene deserts* could be attributed to the unusually CpG rich nature of SINE elements that are

potential targets for genomic methylation (Yoder et al. 1997). These regions could act as methylation nucleation centers and extend this effect out onto the neighboring non-transposable regions (Hasse and Schulz 1994; Rubin et al. 1994). Alu-originated methylation, which is associated with suppression of gene transcription in imprinted regions (Greally 2002), can result in blocking distant gene regulation, by disrupting regulatory elements scattered throughout the *gene deserts*. If this is the case, evolutionarily forces could work against overpopulating *gene deserts* with SINE repetitive elements.

Evolutonarily Conservation Of Human Gene Deserts

We analyzed the distribution of human/mouse (h/m) and human/chicken (h/c) evolutionary conserved regions (ECRs) in human *gene deserts* and regular intergenic regions. While the density of h/m ECRs was found to be 17% higher in regular intergenic regions (intergenic regions: 3.79 ECRs/10kb; *gene deserts*: 3.24 ECRs/10kb), the same ECR density was found in h/c comparisons for both types of intergenic regions (0.33 ECRs /10kb). Interestingly, noncoding ECRs (ncECR) in *gene deserts* are longer than those found in regular intergenic regions, with an average h/m ECR length in *gene deserts* of 265 bps and that in regular intergenic regions of 218 bps. Human and chicken alignments reveal even longer ECRs, with an average h/c ECR of 282 bps for *gene desert* regions, but shorter 203 bps for the regular intergenic intervals. This suggests different evolutionarily behavior for these two types of intergenic intervals. The number of ECRs in regular intergenic intervals appears to be rapidly decreasing over evolutionary time, and they shrink in size as we move from human and mouse to more distantly related organisms. By contrast, in the case of *gene deserts* we observe the preservation of longer

ECRs. This effect is even more evident in *stable gene deserts*; in these intervals, h/m ECRs average 288 bps in length, while h/c ECRs span 304 bps on average. This increased average size of deeply conserved ECRs in stable gene deserts readily suggests a new strategy for the identification of ECRs with higher likelihood to be functional. Additional support for this observation was obtained by computing the sizes of ECRs that have been shown to be functional in multiple studies. As additional support for this observation, it is worth mentioning that known functional h/m ECRs have been shown to be in average larger than 350bps (Ovcharenko et al. 2004, in press).

In order to study distant vertebrate evolution of ncECRs in *gene deserts*, we analyzed the distribution of 2,968 human-fugu (h/f) ncECRs across different regions in the genome. A very similar density of h/f ncECRs was found in *gene deserts* and across the whole genome (10.9 and 10.4 ncECRs/1Mb, correspondingly). Strikingly, we found that the density of h/f ncECRs differed dramatically between *stable* and *flexible gene deserts*. Ninety eight percent of h/f ncECRs (760 out of 777) located in *gene deserts* were found in *stable gene deserts*, even though *stable gene deserts* cover only 29% of the total length of sequences corresponding to *gene deserts*. This brings the density of h/f ncECRs up to 36.7 ncECRs/Mb for *stable gene deserts,* a 3.5-fold increase compared to the average for the whole genome and 122-fold increase compared to the corresponding density in *flexible gene deserts*. This distinct partitioning of these two *gene desert* categories suggests fundamentally different functions for *stable* and *flexible gene deserts*. Most likely, *stable gene deserts* represent the treasure boxes of key distant regulatory elements that are preserved throughout vertebrate evolution. At the same time, *flexible gene deserts*, completely devoid of conservation to fish, and only marginally conserved to chickens, likely have different functions (if any) in the human genome, and likely

specific to higher vertebrates.

We also examined *gene deserts* for the presence of short DNA islands that mutate at significantly lower rates than the average mutation rate in the human genome, termed *ultraconserved* elements (Bejerano et al. 2004). Two thresholds, slightly different from the ones used in the original definition of these regions, have been independently used to identify approximately the same number of h/m and h/c *ultraconserved* elements (approximately, 7,000 elements for both comparison). Only 37% of these two datasets of *ultraconserved* elements overlap, suggesting that *ultraconserved* elements in the mammalian lineage are not necessarily preserved as *ultraconserved* in the more distantly related chicken genome. Interestingly, the density of *ultraconserved* h/c elements is very similar in *gene deserts* compared to the average in the genome, while only half of the average density in the genome in h/m comparisons (Figure 4). Interestingly, the density of *ultraconserved* elements varies greatly between *stable* and *flexible gene deserts*, with most ultraconserved elements located within *stable deserts* and only negligible traces of *ultraconserved* elements identified in *flexible gene deserts, both* in h/m and h/c comparisons. These data suggest that whatever functions are associated with *ultraconserved* elements – they are likely missing in *flexible gene deserts* and are enriched in *stable gene deserts*.

Finally, we observed that the probability for a h/m ncECR to also be conserved in chickens was significantly higher for UTRs than for all other noncoding elements. While only 7.6% of h/m ncECRs were also conserved in chicken, 25.3% of those h/m ncECRs that overlap with 5' or 3' UTRs were conserved in chicken. This approximately 4-fold increase in the ratio of h/m ncECRs that are also conserved with chicken specific to UTRs suggests that an increased selection pressure applies to UTRs with the ECRs

probably highlighting functional regions inside the UTRs. It is also possible that h/c conserved UTRs preferably indicate UTRs of genes with regulatory elements embedded into their untranslated regions (Hillier and al. 2004, submitted). We analyzed the dependence of genes density with h/c UTR ECRs on the density of genes across different human chromosomes (Figure 5), and observed a strong negative correlation between gene density and UTR conservation. For example, the most gene-rich human chromosome 19 had the lowest percentage of genes with conserved UTRs, while gene-poor chromosomes 13 and 18 demonstrated over 55% of genes with UTRs conserved between humans and chickens. The fact that UTRs of genes from gene dense regions are probably depleted in regulatory elements in addition to the previous observation that genes in gene dense regions are also depleted in distant regulatory elements suggests that the regulation of genes in these distinct genomic fractions is fundamentally different. This also suggests that the regulation of genes within gene-rich regions possibly is determined primarily through only promoter and/or intronic regulatory elements.

Stable Gene Deserts Are Linked To Neighboring Genes

The availability of the complete sequence for the chicken genome allowed us to address a very important question about the function of *gene deserts:* do gene deserts harbor functional elements directly associated with one or both their flanking genes (such as gene regulatory elements), or do they contain elements that function independently of the neighboring genes (i.e. chromosome stability regions, matrix attachment sites, noncoding RNA genes, etc), and thus likely play functions other than gene regulation? If indeed gene deserts harbor distant regulatory sequences, this would strongly preclude the accumulation of synteny breakpoints within these *gene deserts* (otherwise, a

chromosomal breakpoint within a *gene desert* would destroy or remove a regulatory elements from the gene it regulates). To address the validity of these assumptions, we analyzed the density of h/c and h/m syntenic breakpoints for different types of genomic intervals. It is important to mention that in order to analyze only large-scale rearrangements and to exclude minor breakpoints that can be associated with evolutionarily microrearrangements, such as pseudogenization through retrotransposition or sequence reshuffling guided by transposable elements, we analyzed only large syntenic intervals - blocks of nucleotide sequence similarity spanning more than 50kb in both species. Synteny coverage of different regions was calculated using synteny to all the available chicken sequence (including "random" chromosomes and the chromosome Un containing unplaced sequences), while the syntenic similarity to unplaced chicken contigs was excluded from the analysis to exclude the synteny breaks associated with non-assembled parts of the chicken genome.

Analyzing this dataset of human-chicken synteny blocks, we found that only 2 human *stable gene deserts* (out of 172 total) contained a synteny breakpoint in them. A detailed analysis of these two synteny breakpoints suggested that they are most likely just artifacts introduced during chicken genome assembly (in both cases two chicken homology regions involved into the generation of a synteny breakpoints were located on the same chicken chromosome and were separated by only 4Mb and 8Mb, respectively). Four other stable human *gene deserts* were not reliably mapped to the chicken genome probably illustrating the limitations of the method used here to define the large syntenic blocks. The remaining 166 human *stable gene deserts* were found to have a single syntenic counterpart in the chicken genome displaying very long stability range associated with these intervals. The synteny region spanned over 80% of the length for

95% of these *stable gene deserts*. This finding suggests that *stable gene deserts* are functionally linked to at least one of the flanking genes and most likely represent accumulations of critical gene regulatory elements that act at a distance. Their location, structural linearity and integrity have been preserved throughout the evolution of vertebrate species. This could indicate that arrays of gene regulatory *cis*-elements are embedded throughout the length of *stable gene deserts* preventing their separation from each other and/or from the gene or genes they regulate.

Dramatic differences in the density of synteny breakpoints were also observed between *stable gene deserts*, *gene-rich regions* and average intergenic regions (Figure 6). Interestingly, the density of synteny breakpoints was very high in *gene-rich regions* in comparisons to the average level in the genome for both h/m and h/c comparisons. One explanation may be that; in sharp contrast to *stable gene deserts*, *gene-rich regions* have possibly evolved as hot spots of chromosomal rearrangements both before and after the primate-rodent radiation. This also suggests that the genes embedded within gene-rich segments are not functionally linked to distant regulatory elements as in stable gene deserts, and thus tolerate recombination events in their vicinities. These data further supports the notion that the regulation of genes flanking *gene deserts* and genes within *gene-rich* regions differ in fundamental ways, where genes flanking gene deserts rely on distant gene regulatory elements, while *gene-rich* regions are predominantly regulated by promoter proximal and/or intronic regulatory sequences.

Identification Of Gene Deserts In The Chicken And Mouse Genomes

Long linear syntenic blocks based on dense clustered ECRs that are spanning over 80% of the length of the human *stable gene deserts* allow for the direct and reliable

mapping of orthologous regions in other species. This method of synteny mapping of *gene deserts* excludes any uncertainty in defining *gene desert* associated with an incomplete catalog of annotated genes in both the chicken and mouse genomes. By requiring the original human and the orthologous mapped *gene deserts* to share boundary ECRs, we defined edge markers and consequently were able to reliably calculate the length for the corresponding *gene deserts* from different species. One hundred forty nine human *stable gene deserts* were reliably mapped to the mouse and chicken genomes and identified as a single contiguous sequence stretch in all three species. Using this dataset of h/m and h/c orthologous *stable gene desert* intervals we compared their lengths in three genomes (Figure 7). No significant size differences were observed between human and mouse *gene deserts* beyond differences associated with minor mouse genome shrinkage. Also, the lengths of individual *gene deserts* were highly similar beyond the primate-rodent radiation. A high correlation in lengths was also observed between human and chicken *stable gene deserts* (R^2=0.71). The majority of the analyzed chicken *gene deserts* counterparts consistently were 0.39x the size of their human counterparts, with a few outliers that were either larger or smaller than the average. This is indeed a very interesting observation taking into account that the human-chicken genome expansion rate varies significantly (approximately 5-fold in the magnitude) in the different regions in the human genome. Basically, while there is high flexibility in the short-range rate of genome expansion between humans and chicken (Hillier and al. 2004, submitted), it is not the case for *stable gene deserts*. This could indicate that these regions most likely are resistant to large-scale deletions or rearrangements. The h/c expansion coefficient for *stable gene deserts* is also very close to the average for these genomes [chicken genome size is 0.37 of the human genome if unplaced contigs are

excluded (i.e. chrUn and random pieces)]. This suggests that effects in mammalian genomes such as inflation of repetitive elements have approximately the same rate of appearance in *stable gene deserts* and other genomic intervals. One conclusion from these observations is that the inter-ECR distances in *stable gene deserts* are very elastic and that the putative functions of pairs of ECRs will not be disrupted upon inserting a repetitive element in between. That also suggests that distant regulatory elements in *stable gene deserts* function independently of the distance to the transcriptional start site of the genes they regulate.

A very interesting and unique feature of the chicken genome is an abundance of microchromosomes (varying in size from 1.0 to 20.6 Mb). *A priori,* the small size of these chromosomes would suggest that they are depleted of extensively long *gene deserts*, especially when considering the possibility that microchromosomes may have evolved through multiple rearrangement events, while *stable gene deserts* maintain their structural integrity and lack chromosomal breaks. Contrary to this hypothesis, we did not observe a decrease in the density or size of *stable gene deserts* on microchromosomes (Figure 7, 8), rather the density of *stable gene deserts* was slightly higher in microcromosomes than in all other chromosomal categories. This distribution of *stable gene deserts* in the chicken genome points to the independence of the presence and density of stable gene deserts on the size of individual chromosomes. Also, the level of coverage of microchromosomes by *stable gene deserts* suggests that *stable gene deserts* do not have an obvious bias against appearance of synteny breaks in the surrounding regions. Therefore the stringent framework of chromosomal stability observed within stable gene deserts abruptly disappears immediately beyond the boundaries of *the deserts* and their associated genes. These data indicate that the integrity of stable gene deserts is

primarily dictated by the linear relationship between the genes flanking these deserts and their long-range regulatory elements, and not by other unknown chromosomal architectural or genomic properties.

DISCUSSION

Gene deserts – large intergenic regions that collectively cover 25% of the human genome hold very distinct evolutionarily and sequence signatures that clearly set them apart from the rest of the genome. The GC content, repeat content, chicken and mouse conservation of human *gene deserts* significantly differ from *gene-rich* regions and other regular intergenic regions – two other representative fractions of the human genome. *Gene deserts* accumulate more SNPs and are overall less highly conserved across species. However, these regions preserve their repeat content suggesting that rapid evolutionary changes in these regions with elevated levels of both accumulation and deleterious processes.

Comparative sequence analysis of the human *gene deserts* and their chicken orthologs effectively separates *gene deserts* into two categories – *stable* and *flexible gene deserts*. *Stable gene deserts* display high levels of sequence similarity in humans and chicken, while the *flexible deserts* represent regions specific to the mammalian lineage. *Stable gene deserts* display lower repeat density and as high levels of h/m sequence similarity as the conserved *gene-rich* regions of the genome, suggestive of considerable degrees of purifying pressure acting over these *stable gene deserts*. Moreover, 33% of the *stable gene deserts* cluster in pairs surrounding a small number of well-localized genes creating long and well-shaped islands of genomic sequence with minimum gene density that are much more effectively preserved throughout the evolution of vertebrates than the rest of the genome. Not surprisingly the majority of genes that are either flanked by *stable gene deserts* or are neighboring these highly conserved intervals are functionally related to core biochemical processes of vertebrates such as regulation of transcription, skeletal and muscle development, DNA binding, and regulation of

metabolism.

The last ~100MYs of vertebrate genome evolution is highlighted by an explosive appearance of multiple repetitive elements. Different types of repetitive elements are not uniformly represented in the human genome and we find that gene deserts are enriched in LINE elements while regular intergenic regions have preferably accumulated SINE elements. A profound depletion in SINE elements in gene deserts can potentially be related to SINE-mediated genome methylation, a process responsible for gene silencing that most likely is detrimental to the function of *stable gene deserts* enriched in transcriptional regulatory elements.

The density of h/f ncECRs is negligibly small across *flexible gene deserts* and is simultaneously strongly elevated in *stable gene deserts* separating the biological function and evolutionarily importance of these two categories of *gene deserts*. *Stable gene deserts* that extensively harbor and safeguard noncoding elements throughout evolution of vertebrates are the best candidates for regions with key distant gene regulatory elements in the human genome. The function of *flexible gene deserts* is more ambiguous. They possibly represent recently evolved regions that have not yet been fixed or they may lack important function and represent junkyards in the genome. This potentially reconciles the apparent disparity reported that some of *gene deserts* in the human genome were found to be rich in gene regulatory elements (Nobrega et al. 2003) while others have no phenotypic impact when removed from the mouse genome (Pennisi 2004). By comparing the properties of ECRs in *stable* and *flexible gene deserts*, and regular intergenic intervals, we found that the ECRs in *stable gene deserts* are much longer than the average in the genome and this stands true for both h/m and h/c conservation profiles. The same enrichment associated with *stable gene deserts* was also observed for

ultraconserved h/m and h/c elements. Previously, it was suggested that long ECRs could be functional important emphasizing the functional importance of noncoding elements populating *stable gene deserts*.

The distribution of long-range syntenic blocks interconnecting human, chicken, and mouse chromosomes as overlapped with the distribution of *gene-rich* regions and *stable gene deserts* revealed distinct structural evolutionary events unique to each one of these genomic intervals. While we found that *gene-rich* regions accumulate synteny breakpoints twice as fast as the average intergenic regions, *stable gene deserts* were completely depleted of synteny breakpoints. Ninety six percent of *stable gene deserts* are represented as a single syntenic block in the genomes of these three species despite the extremely large size of these genomic intervals. The almost absolute preservation of chromosomal integrity of *gene deserts* suggests that the regulation of genes flanking *gene deserts* and that of genes contained within *gene-rich* regions differs. Genes bracketing *stable gene deserts* most probably have distant gene regulatory elements that cannot be separated by recombination events, while the regulation of the genes within *gene-rich* genomic regions takes places through promoters, and intronic sequences, or, less probable, through UTR elements. The negative correlation in UTR conservation with gene density suggests that there are many genes that have functional elements in the UTR regions, but those genes probably are not associated with either gene deserts or gene rich regions.

By using contiguous synteny relationships for the human genome with the genomes of mice and chicken in *stable gene deserts* regions we were able to identify *stable gene deserts* in chicken and mice without requiring a reliable gene annotation for these two genomes. Analysis of the length difference of *stable gene deserts* in different

species pointed to the absence of large scale genomic events in *stable gene deserts* as highlighted by a very similar length of the human stable gene deserts and their mouse counterparts as well as by the length difference of all the human and chicken individual *gene deserts* that strongly correlates with the human genome expansion coefficient. The uniform expansion of individual human *stable gene deserts* implies that the function of distant regulatory elements is independent of the distance between neighboring regulatory elements or the regulatory elements and the corresponding genes providing some insights on the distant regulatory activity. Also, the distribution of chicken *stable gene deserts* in the chicken genome is not diminished in microchromosomes suggesting that gene deserts-associated chromosomal stability abruptly disappears beyond the boundaries of the *gene deserts* and their associated genes. Our evolutionary analysis emphasizes on the importance of *stable gene deserts* and suggests that are likely to play a critical biological role in vertebrates.

METHODS

Identification of ECRs and ultraconserved regions.

The analysis of syntenic relationships and conservation profiles was done through the annotation of evolutionarily conserved regions (ECRs) in the alignments of genomes. We employed the genome alignments generated by the ECR Browser (http://ecrbrowser.dcode.org) (Ovcharenko et al. 2004). A genomic interval was annotated as an ECR if it was >100 bps and >70% identity as defined by the number of nucleotide matches in a sliding window. 184k ECRs were identified in h/c alignments and 1,268k ECRs in h/m alignments. 16% h/m and 59% h/c ECRs overlapped with exons of 'known' genes presenting a significant imbalance of h/c nucleotide conservation

in protein coding regions.

A scan for *ultraconserved* regions (Bejerano et al. 2004) was performed with different parameters for the h/m and h/c alignments. Consistent with previous reports a threshold of >200 bps/99% identity was used to identify 6,849 *ultraconserved* regions in h/m alignments (we increased the flexibility of the definition to account for putative minor changes and sequencing errors). 6,677 *ultraconserved* elements in h/c alignments were identified using a >200 bps/95% criteria. 23% of chicken and 36% of mouse *ultraconserved* elements overlapped with exonic sequences.

Sixty six thousand h/f ncECRs (>100bp/70%ID) were identified as described (Ovcharenko et al., 2004; in press). A deeper filtering out known and putative transcripts, pseudogenes, mRNAs, as well as proximal promoter sequences resulted in 2,968 h/f ncECRs that do now possess any protein coding activity and are distantly positioned from the transcriptional start site of adjacent genes.

Defining SINE-ages.

SINE-age of a genomic region *X* is defined based on ratios of coverage of the region by different SINE families:

$$SA_X = \sum_{i \subset SINE} c_i \cdot age_i ,$$

where the summation is performed over all the different age-defining SINE families (AluY, AluS*, AluJ*, and FAM). Also, age_i, and c_i denote the age and the content ratio SINE family *i* populating a particular region ($\sum_{i \subset SINE} c_i = 1$). Relative SINE-age of a region is defined by a subtraction of the average SINE-age of the genome from the original SINE-age of the region:

$$\overline{SA}_X = SA_X - SA_{genome}$$

This equation quantifies an increase in the density of younger Alu repetitive elements by a positive value and a decrease by a negative value. The absolute value of relative SINE-age presents the level of bias from the average genomic distribution. The result is in MYs and SA_{genome} was found to be equal to 46.0 MYs.

Large blocks of synteny.

In order to create a map of genomes synteny, which is based on nucleotide-type alignments, we scanned the dataset of all the triplets of ECRs consecutively present in both species (two neighboring ECRs were selected as consecutively located only if they were separated by <100kb in both genomes). Constructed ECR triplets defined anchors of genome similarity and were used to construct long syntenic blocks by clustering ECR triplets together using the same 100kb threshold again. A filtering out of regions that cover less than 50kb in one of the species created a dataset of long regions of synteny. Subsequently the joining of these long regions of synteny was performed into longer regions of synteny if the separation of a pair of long regions of synteny were shorter than 1Mb in both genomes. Taking into account that we used a single ECR coverage of every genome, the identified synteny similarity of the genomes was approximately orthologous (tandem gene duplication events were incorporated into the blocks of long orthology).

Using this approach large scale similarity of the human and mouse genomes was modeled with ca 300 and 500 synteny breakpoints between human and mouse and human and chicken genomes, correspondingly (chicken chromosomes Un, random, and several others representing unassembled chicken sequence were excluded from consideration). Due to longer evolutionarily separation of birds from humans comparing with the

separation of rodents from humans, we observed different levels of genome coverage by the syntenic blocks in human-mouse and human-chicken comparisons. Approximately 96% of mammalian genomes were covered by h/m large syntenic blocks. H/c large syntenic blocks covered 90% of the chicken genome and 78% of the human genome.

ACKNOWLEDGEMENTS

W.M. and R. H. were supported by NHGRI grant HG02238; G.G.L. was supported by LLNL LDRD-04-ERD-052 grant; I.O. was in part supported by DOE SCW0345 grant. The work was performed under the auspices of the U.S. Department of Energy by the University of California, Lawrence Livermore National Laboratory Contract No. W-7405-Eng-48.

Region	Length, Mb	GC content	Chicken conservation*	Mouse conservation*	Repeat content	Density of SNPs
Deserts	716	37.45%	1.91%	19.04%	50.50%	0.73/kb
Regular	56	46.14%	1.41%	18.94%	51.85%	0.55/kb
Gene-rich	285	47.35%	4.35%	27.98%	48.87%	0.57/kb
Average	2,842	40.87%	2.98%	22.38%	48.54%	0.66/kb

Table 1. Characteristic features of gene deserts, gene-rich regions, regular intergenic regions, and the average in the human genome, NCBI Build 34. Repeat content and SNP annotation derived from the tabular genome annotation obtained from the UCSC Genome Browser utility. * Interspecies conservation describes the percentage of non-repetitive sequence covered by the ECRs.

Category	Enrichment	Classification
stable gene deserts		
regulation of metabolism	4.4	biological process
transcription factor activity	4.2	molecular function
transcription coactivator activity	4.0	molecular function
regulation of biosynthesis	3.8	biological process
transcription regulator activity	3.6	molecular function
transcription factor binding	3.2	molecular function
DNA binding	2.8	molecular function
regulation of transcription	2.8	biological process
transcription	2.7	biological process
development	2.0	biological process
flexible gene deserts		
glutamate receptor activity	7.8	molecular function
inotropic glutamate receptor activity	7.7	molecular function
amine receptor activity	6.2	molecular function
sulfotransferase activity	4.2	molecular function
cell adhesion	3.0	biological process
transmission of nerve impulse	2.8	biological process
neuromuscular physiological process	2.8	biological process
synaptic transmission	2.7	biological process
calcium ion binding	2.2	molecular function
organogenesis	1.9	biological process
morphogenesis	1.7	biological process
development	1.7	biological process
cell communication	1.6	biological process

Table 2. Enrichment in Gene Ontology categories for *stable* and *flexible gene deserts* (the statistical significance of the reported numbers is supported by the p-values $<1e^{-5}$ as quantified in a comparison with the purely-by-chance expectations).

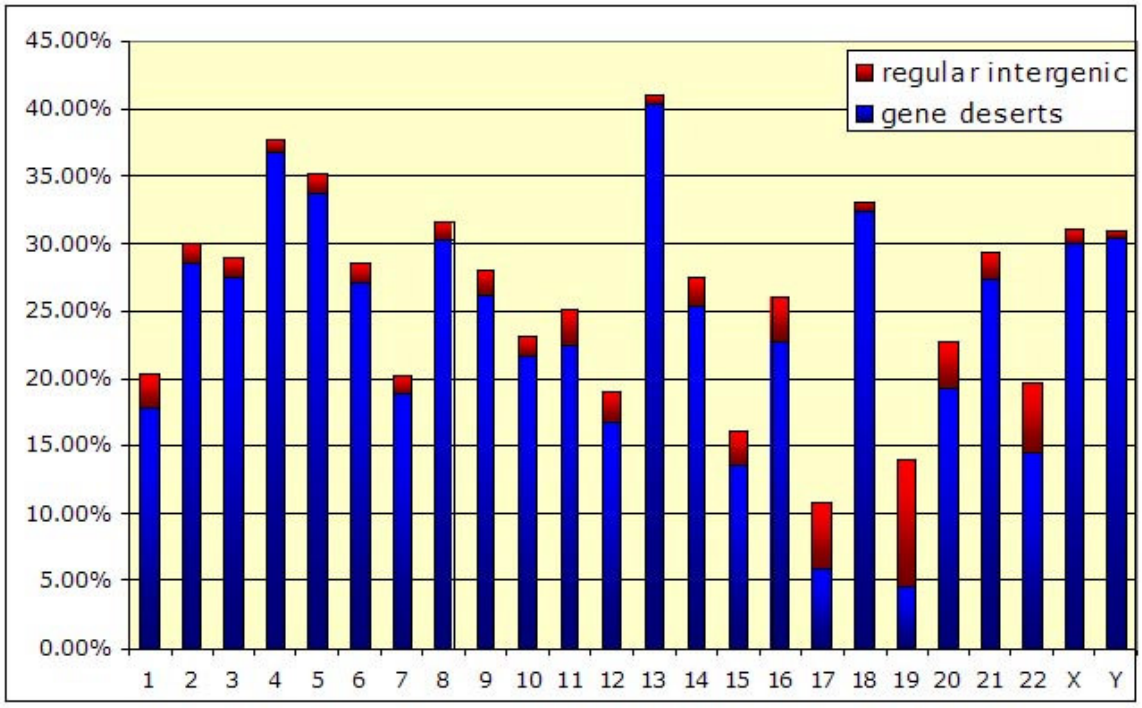

Figure 1. Chromosome coverage by gene deserts (in blue) and regular intergenic regions (in red).

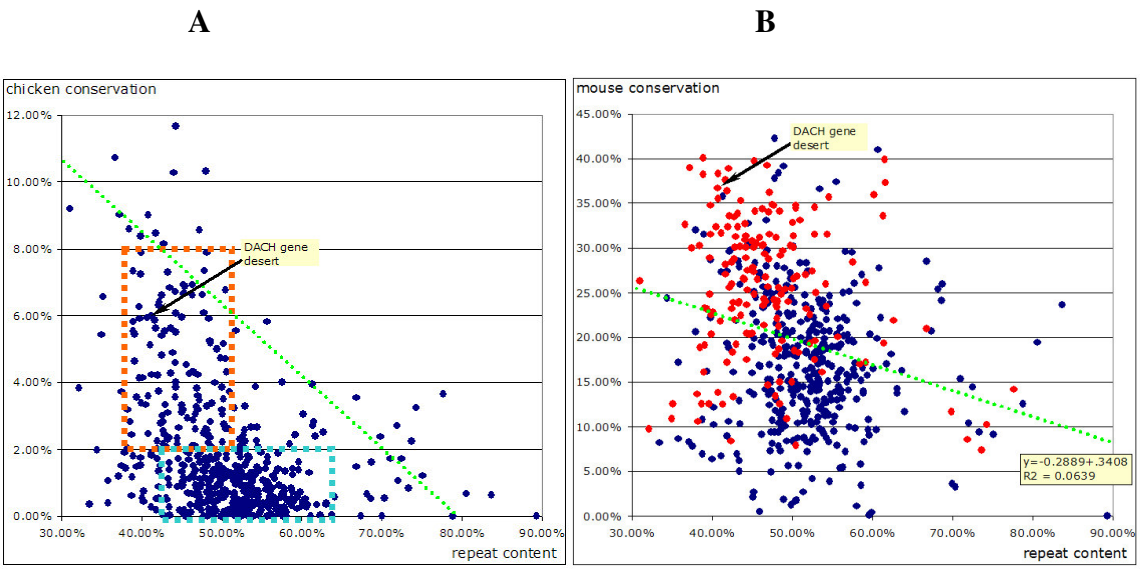

Figure 2. Correlation of conservation of the human gene deserts with chicken (A) and mouse (B) vs. repeat content. Conservation length is assessed as the total length of the sequence underlying ECRs in a region. Three distinct groups of gene deserts can be identified through the comparisons of human and chicken sequences: poorly conserved (blue rectangle); well conserved with low variability in repeat content (orange rectangle); and a limited number of outliers in repeat content and conservation that show an almost linear negative correlation of conservation level and repeat content (surrounding green line). Negative correlation of the mouse conservation level and repeat content is very weak (R^2=0.06). Stable gene deserts that are well conserved with chickens (>=2% level of non-repetitive conservation; depicted in red) show in general a higher level of human-mouse conservation then the flexible gene deserts that are poorly conserved in chicken (<2% level of non-repetitive conservation; in blue).

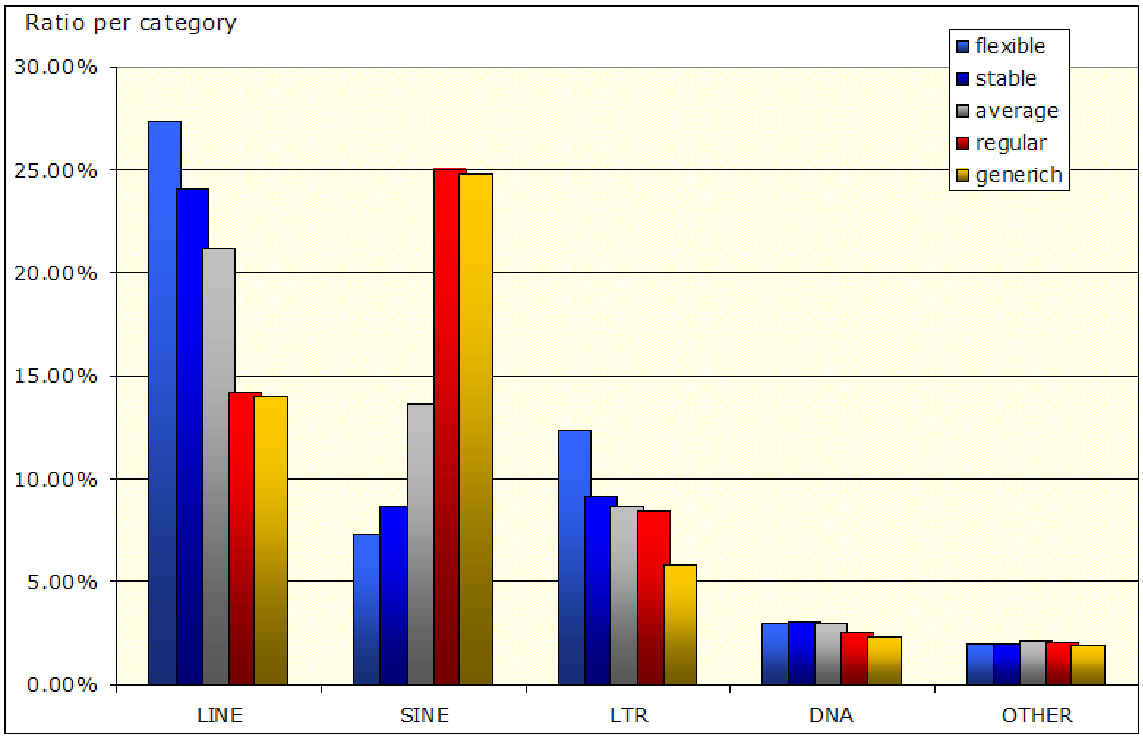

Figure 3. Ratio of different categories of repetitive elements populating different genomic regions. Flexible gene deserts are in light blue, stable gene deserts are in blue, average counts for the human genome are in gray, regular intergenic regions are in red, and gene-rich regions are in yellow.

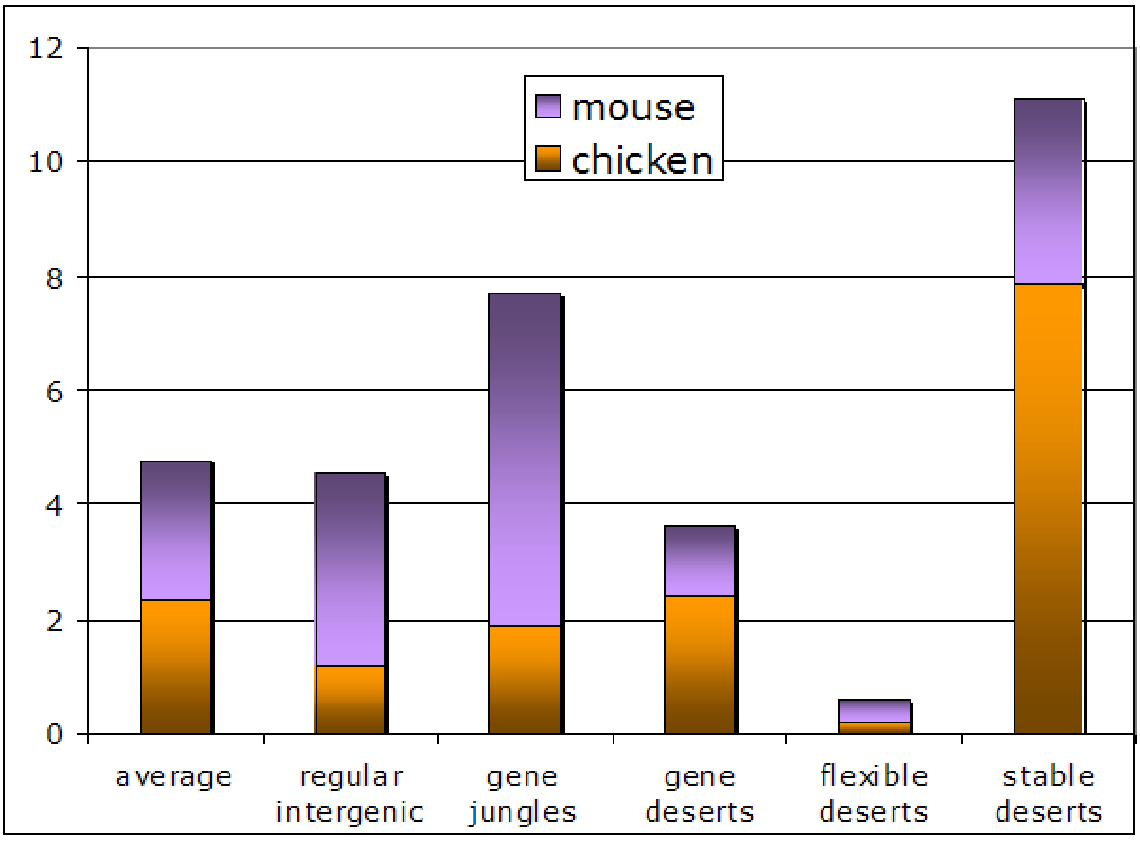

Figure 4. Density of human-mouse (in lilac) and human-chicken (in orange) "ultraconserved" elements in different regions. Vertical axis is scaled as a number of "ultraconserved" elements per 1 Mb of sequence.

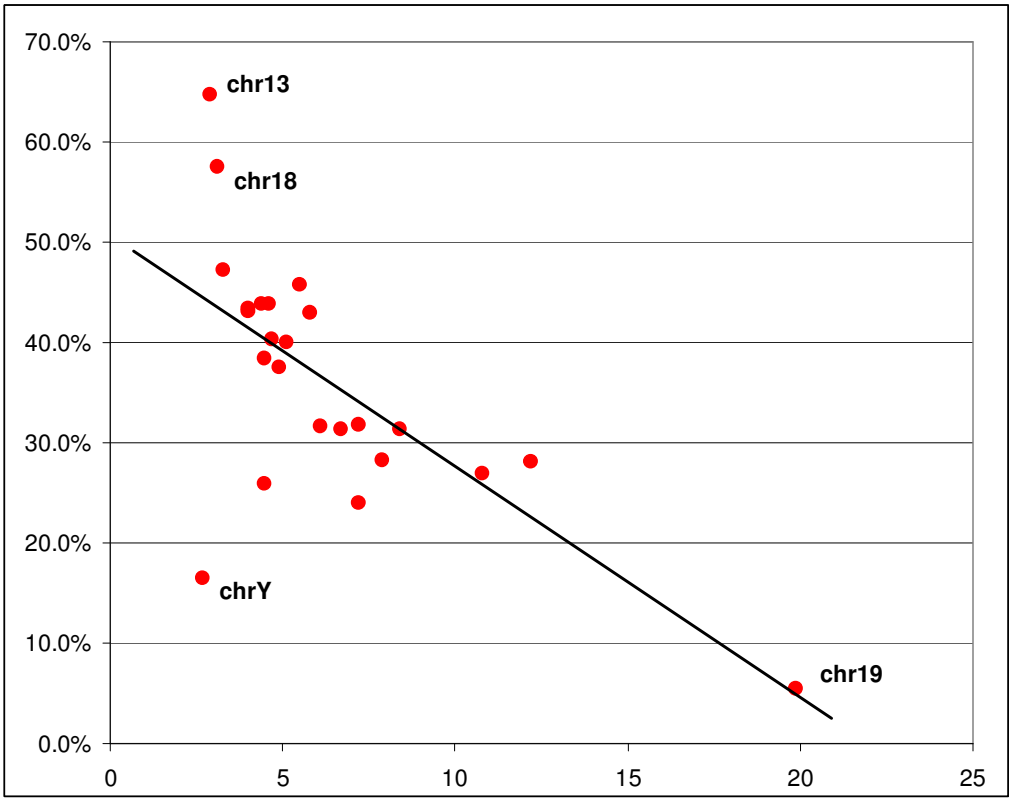

Figure 5. Percentage of the gene with UTRs conserved in chicken (vertical axis) versus the gene density (based on RefSeq annotation; in genes per 1Mb of sequence as plotted at the horizontal axis). Red dots describe different human chromosomes.

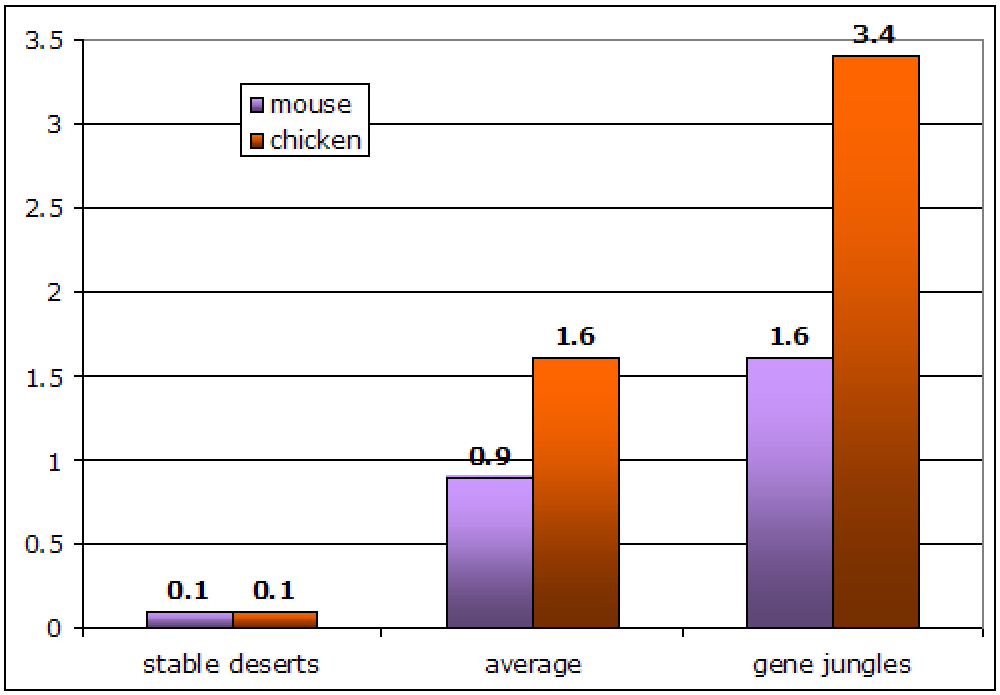

Figure 6. Density of synteny breakpoints per 1 Mb of sequence. Human-mouse comparisons are in orange, human-chicken in lilac.

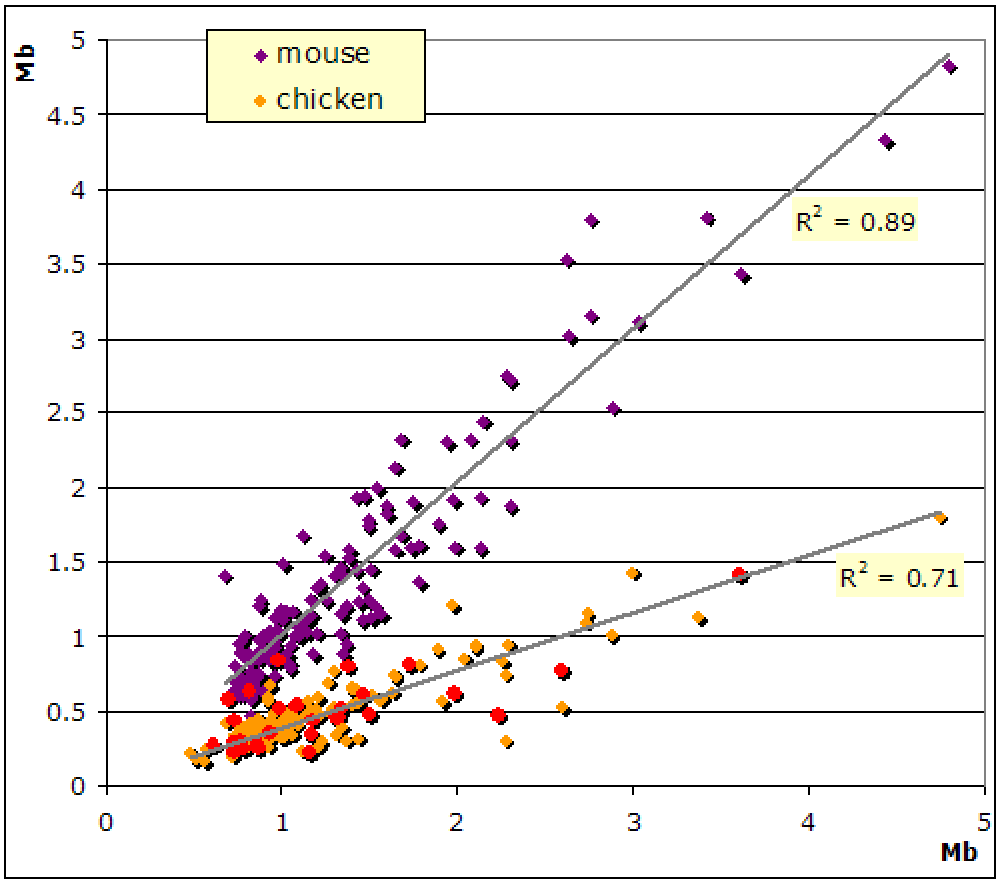

Figure 7. Length of orthologous *stable gene desert* counterparts in the chicken and mouse genomes as compared to the human genome. *Gene deserts* from chicken microchromosomes are in red.

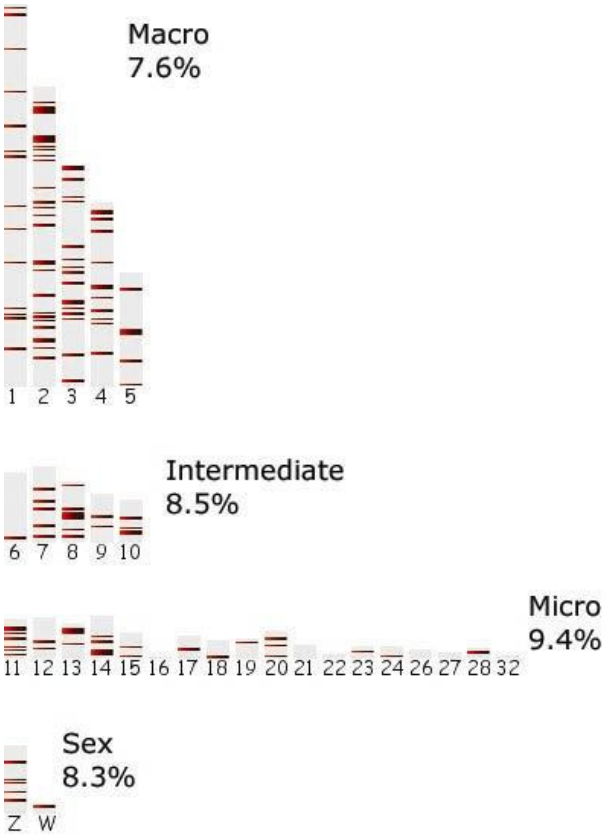

Figure 8. Distribution of stable gene deserts in the chicken genome (plotted as red lines). Chicken chromosomes are grouped into “Macro”, “Intermediate”, “Micro”, and “Sex” categories with the numerical characterization of average chromosome coverage by the stable gene deserts.

REFERENCES

Bejerano, G., M. Pheasant, I. Makunin, S. Stephen, W.J. Kent, J.S. Mattick, and D. Haussler. 2004. Ultraconserved elements in the human genome. *Science* **304:** 1321-1325.

Dehal, P., P. Predki, A.S. Olsen, A. Kobayashi, P. Folta, S. Lucas, M. Land, A. Terry, C.L. Ecale Zhou, S. Rash, Q. Zhang, L. Gordon, J. Kim, C. Elkin, M.J. Pollard, P. Richardson, D. Rokhsar, E. Uberbacher, T. Hawkins, E. Branscomb, and L. Stubbs. 2001. Human chromosome 19 and related regions in mouse: conservative and lineage-specific evolution. *Science* **293:** 104-111.

Dunham, A. L.H. Matthews J. Burton J.L. Ashurst K.L. Howe K.J. Ashcroft D.M. Beare D.C. Burford S.E. Hunt S. Griffiths-Jones M.C. Jones S.J. Keenan K. Oliver C.E. Scott R. Ainscough J.P. Almeida K.D. Ambrose D.T. Andrews R.I. Ashwell A.K. Babbage C.L. Bagguley J. Bailey R. Bannerjee K.F. Barlow K. Bates H. Beasley C.P. Bird S. Bray-Allen A.J. Brown J.Y. Brown W. Burrill C. Carder N.P. Carter J.C. Chapman M.E. Clamp S.Y. Clark G. Clarke C.M. Clee S.C. Clegg V. Cobley J.E. Collins N. Corby G.J. Coville P. Deloukas P. Dhami I. Dunham M. Dunn M.E. Earthrowl A.G. Ellington L. Faulkner A.G. Frankish J. Frankland L. French P. Garner J. Garnett J.G. Gilbert C.J. Gilson J. Ghori D.V. Grafham S.M. Gribble C. Griffiths R.E. Hall S. Hammond J.L. Harley E.A. Hart P.D. Heath P.J. Howden E.J. Huckle P.J. Hunt A.R. Hunt C. Johnson D. Johnson M. Kay A.M. Kimberley A. King G.K. Laird C.J. Langford S. Lawlor D.A. Leongamornlert D.M. Lloyd C. Lloyd J.E. Loveland J. Lovell S. Martin M. Mashreghi-Mohammadi S.J. McLaren A. McMurray S. Milne M.J. Moore T. Nickerson S.A. Palmer A.V. Pearce A.I. Peck S. Pelan B. Phillimore K.M. Porter C.M. Rice S. Searle H.K. Sehra R. Shownkeen C.D. Skuce M. Smith C.A. Steward N. Sycamore J. Tester D.W. Thomas A. Tracey A. Tromans B. Tubby M. Wall J.M. Wallis A.P. West S.L. Whitehead D.L. Willey L. Wilming P.W. Wray M.W. Wright L. Young A. Coulson R. Durbin T. Hubbard J.E. Sulston S. Beck D.R. Bentley J. Rogers and M.T. Ross. 2004. The DNA sequence and analysis of human chromosome 13. *Nature* **428:** 522-528.

Greally, J.M. 2002. Short interspersed transposable elements (SINEs) are excluded from imprinted regions in the human genome. *Proc Natl Acad Sci U S A* **99:** 327-332.

Grimwood, J., L.A. Gordon, A. Olsen, A. Terry, J. Schmutz, J. Lamerdin, U. Hellsten, D. Goodstein, O. Couronne, M. Tran-Gyamfi, A. Aerts, M. Altherr, L. Ashworth, E. Bajorek, S. Black, E. Branscomb, S. Caenepeel, A. Carrano, C. Caoile, Y.M. Chan, M. Christensen, C.A. Cleland, A. Copeland, E. Dalin, P. Dehal, M. Denys, J.C. Detter, J. Escobar, D. Flowers, D. Fotopulos, C. Garcia, A.M. Georgescu, T. Glavina, M. Gomez, E. Gonzales, M. Groza, N. Hammon, T. Hawkins, L. Haydu, I. Ho, W. Huang, S. Israni, J. Jett, K. Kadner, H. Kimball, A. Kobayashi, V. Larionov, S.H. Leem, F. Lopez, Y. Lou, S. Lowry, S. Malfatti, D. Martinez, P. McCready, C. Medina, J. Morgan, K. Nelson, M. Nolan, I. Ovcharenko, S. Pitluck, M. Pollard, A.P. Popkie, P. Predki, G. Quan, L. Ramirez, S. Rash, J. Retterer, A. Rodriguez, S. Rogers, A. Salamov, A. Salazar, X. She, D. Smith, T. Slezak, V. Solovyev, N. Thayer, H. Tice, M. Tsai, A. Ustaszewska, N. Vo, M. Wagner, J. Wheeler, K. Wu, G. Xie, J. Yang, I. Dubchak, T.S. Furey, P. DeJong, M. Dickson, D. Gordon, E.E. Eichler, L.A. Pennacchio, P. Richardson, L. Stubbs, D.S. Rokhsar, R.M. Myers, E.M. Rubin, and S.M. Lucas. 2004. The DNA

sequence and biology of human chromosome 19. *Nature* **428:** 529-535.
Hasse, A. and W.A. Schulz. 1994. Enhancement of reporter gene de novo methylation by DNA fragments from the alpha-fetoprotein control region. *J Biol Chem* **269:** 1821-1826.
Hillier, L.W. and e. al. 2004, submitted. Sequencing and comparative analysis of the chicken genome. *Nature*.
Karolchik, D., R. Baertsch, M. Diekhans, T.S. Furey, A. Hinrichs, Y.T. Lu, K.M. Roskin, M. Schwartz, C.W. Sugnet, D.J. Thomas, R.J. Weber, D. Haussler, and W.J. Kent. 2003. The UCSC Genome Browser Database. *Nucleic Acids Res* **31:** 51-54.
Lander, E.S. L.M. Linton B. Birren C. Nusbaum M.C. Zody J. Baldwin K. Devon K. Dewar M. Doyle W. FitzHugh R. Funke D. Gage K. Harris A. Heaford J. Howland L. Kann J. Lehoczky R. LeVine P. McEwan K. McKernan J. Meldrim J.P. Mesirov C. Miranda W. Morris J. Naylor C. Raymond M. Rosetti R. Santos A. Sheridan C. Sougnez N. Stange-Thomann N. Stojanovic A. Subramanian D. Wyman J. Rogers J. Sulston R. Ainscough S. Beck D. Bentley J. Burton C. Clee N. Carter A. Coulson R. Deadman P. Deloukas A. Dunham I. Dunham R. Durbin L. French D. Grafham S. Gregory T. Hubbard S. Humphray A. Hunt M. Jones C. Lloyd A. McMurray L. Matthews S. Mercer S. Milne J.C. Mullikin A. Mungall R. Plumb M. Ross R. Shownkeen S. Sims R.H. Waterston R.K. Wilson L.W. Hillier J.D. McPherson M.A. Marra E.R. Mardis L.A. Fulton A.T. Chinwalla K.H. Pepin W.R. Gish S.L. Chissoe M.C. Wendl K.D. Delehaunty T.L. Miner A. Delehaunty J.B. Kramer L.L. Cook R.S. Fulton D.L. Johnson P.J. Minx S.W. Clifton T. Hawkins E. Branscomb P. Predki P. Richardson S. Wenning T. Slezak N. Doggett J.F. Cheng A. Olsen S. Lucas C. Elkin E. Uberbacher M. Frazier R.A. Gibbs D.M. Muzny S.E. Scherer J.B. Bouck E.J. Sodergren K.C. Worley C.M. Rives J.H. Gorrell M.L. Metzker S.L. Naylor R.S. Kucherlapati D.L. Nelson G.M. Weinstock Y. Sakaki A. Fujiyama M. Hattori T. Yada A. Toyoda T. Itoh C. Kawagoe H. Watanabe Y. Totoki T. Taylor J. Weissenbach R. Heilig W. Saurin F. Artiguenave P. Brottier T. Bruls E. Pelletier C. Robert P. Wincker D.R. Smith L. Doucette-Stamm M. Rubenfield K. Weinstock H.M. Lee J. Dubois A. Rosenthal M. Platzer G. Nyakatura S. Taudien A. Rump H. Yang J. Yu J. Wang G. Huang J. Gu L. Hood L. Rowen A. Madan S. Qin R.W. Davis N.A. Federspiel A.P. Abola M.J. Proctor R.M. Myers J. Schmutz M. Dickson J. Grimwood D.R. Cox M.V. Olson R. Kaul N. Shimizu K. Kawasaki S. Minoshima G.A. Evans M. Athanasiou R. Schultz B.A. Roe F. Chen H. Pan J. Ramser H. Lehrach R. Reinhardt W.R. McCombie M. de la Bastide N. Dedhia H. Blocker K. Hornischer G. Nordsiek R. Agarwala L. Aravind J.A. Bailey A. Bateman S. Batzoglou E. Birney P. Bork D.G. Brown C.B. Burge L. Cerutti H.C. Chen D. Church M. Clamp R.R. Copley T. Doerks S.R. Eddy E.E. Eichler T.S. Furey J. Galagan J.G. Gilbert C. Harmon Y. Hayashizaki D. Haussler H. Hermjakob K. Hokamp W. Jang L.S. Johnson T.A. Jones S. Kasif A. Kaspryzk S. Kennedy W.J. Kent P. Kitts E.V. Koonin I. Korf D. Kulp D. Lancet T.M. Lowe A. McLysaght T. Mikkelsen J.V. Moran N. Mulder V.J. Pollara C.P. Ponting G. Schuler J. Schultz G. Slater A.F. Smit E. Stupka J. Szustakowski D. Thierry-Mieg J. Thierry-Mieg L. Wagner J. Wallis R. Wheeler A. Williams Y.I. Wolf K.H. Wolfe S.P. Yang R.F. Yeh F. Collins M.S. Guyer J. Peterson A. Felsenfeld K.A. Wetterstrand A. Patrinos M.J. Morgan J. Szustakowki P. de Jong J.J. Catanese K. Osoegawa H. Shizuya S. Choi and Y.J. Chen. 2001. Initial sequencing and analysis of the

human genome. *Nature* **409:** 860-921.
Nelson, C.E., B.M. Hersh, and S.B. Carroll. 2004. The regulatory content of intergenic DNA shapes genome architecture. *Genome Biol* **5:** R25.
Nobrega, M.A., I. Ovcharenko, V. Afzal, and E.M. Rubin. 2003. Scanning human gene deserts for long-range enhancers. *Science* **302:** 413.
Ovcharenko, I., M.A. Nobrega, G.G. Loots, and L. Stubbs. 2004. ECR Browser: a tool for visualizing and accessing data from comparisons of multiple vertebrate genomes. *Nucleic Acids Res* **32:** W280-W286.
Ovcharenko, I., L. Stubbs, and G.G. Loots. 2004, in press. Interpreting Mammalian Evolution using Fugu Genome Comparisons. *Genomics*.
Pennisi, E. 2004. The Biology of Genomes meeting. Disposable DNA puzzles researchers. *Science* **304:** 1590-1591.
Rubin, C.M., C.A. VandeVoort, R.L. Teplitz, and C.W. Schmid. 1994. Alu repeated DNAs are differentially methylated in primate germ cells. *Nucleic Acids Res* **22:** 5121-5127.
Shannon, M., A.T. Hamilton, L. Gordon, E. Branscomb, and L. Stubbs. 2003. Differential expansion of zinc-finger transcription factor loci in homologous human and mouse gene clusters. *Genome Res* **13:** 1097-1110.
Venter, J.C. M.D. Adams E.W. Myers P.W. Li R.J. Mural G.G. Sutton H.O. Smith M. Yandell C.A. Evans R.A. Holt J.D. Gocayne P. Amanatides R.M. Ballew D.H. Huson J.R. Wortman Q. Zhang C.D. Kodira X.H. Zheng L. Chen M. Skupski G. Subramanian P.D. Thomas J. Zhang G.L. Gabor Miklos C. Nelson S. Broder A.G. Clark J. Nadeau V.A. McKusick N. Zinder A.J. Levine R.J. Roberts M. Simon C. Slayman M. Hunkapiller R. Bolanos A. Delcher I. Dew D. Fasulo M. Flanigan L. Florea A. Halpern S. Hannenhalli S. Kravitz S. Levy C. Mobarry K. Reinert K. Remington J. Abu-Threideh E. Beasley K. Biddick V. Bonazzi R. Brandon M. Cargill I. Chandramouliswaran R. Charlab K. Chaturvedi Z. Deng V. Di Francesco P. Dunn K. Eilbeck C. Evangelista A.E. Gabrielian W. Gan W. Ge F. Gong Z. Gu P. Guan T.J. Heiman M.E. Higgins R.R. Ji Z. Ke K.A. Ketchum Z. Lai Y. Lei Z. Li J. Li Y. Liang X. Lin F. Lu G.V. Merkulov N. Milshina H.M. Moore A.K. Naik V.A. Narayan B. Neelam D. Nusskern D.B. Rusch S. Salzberg W. Shao B. Shue J. Sun Z. Wang A. Wang X. Wang J. Wang M. Wei R. Wides C. Xiao C. Yan A. Yao J. Ye M. Zhan W. Zhang H. Zhang Q. Zhao L. Zheng F. Zhong W. Zhong S. Zhu S. Zhao D. Gilbert S. Baumhueter G. Spier C. Carter A. Cravchik T. Woodage F. Ali H. An A. Awe D. Baldwin H. Baden M. Barnstead I. Barrow K. Beeson D. Busam A. Carver A. Center M.L. Cheng L. Curry S. Danaher L. Davenport R. Desilets S. Dietz K. Dodson L. Doup S. Ferriera N. Garg A. Gluecksmann B. Hart J. Haynes C. Haynes C. Heiner S. Hladun D. Hostin J. Houck T. Howland C. Ibegwam J. Johnson F. Kalush L. Kline S. Koduru A. Love F. Mann D. May S. McCawley T. McIntosh I. McMullen M. Moy L. Moy B. Murphy K. Nelson C. Pfannkoch E. Pratts V. Puri H. Qureshi M. Reardon R. Rodriguez Y.H. Rogers D. Romblad B. Ruhfel R. Scott C. Sitter M. Smallwood E. Stewart R. Strong E. Suh R. Thomas N.N. Tint S. Tse C. Vech G. Wang J. Wetter S. Williams M. Williams S. Windsor E. Winn-Deen K. Wolfe J. Zaveri K. Zaveri J.F. Abril R. Guigo M.J. Campbell K.V. Sjolander B. Karlak A. Kejariwal H. Mi B. Lazareva T. Hatton A. Narechania K. Diemer A. Muruganujan N. Guo S. Sato V. Bafna S. Istrail R. Lippert R. Schwartz B. Walenz S. Yooseph D. Allen A. Basu J. Baxendale L. Blick M. Caminha J.

Carnes-Stine P. Caulk Y.H. Chiang M. Coyne C. Dahlke A. Mays M. Dombroski M. Donnelly D. Ely S. Esparham C. Fosler H. Gire S. Glanowski K. Glasser A. Glodek M. Gorokhov K. Graham B. Gropman M. Harris J. Heil S. Henderson J. Hoover D. Jennings C. Jordan J. Jordan J. Kasha L. Kagan C. Kraft A. Levitsky M. Lewis X. Liu J. Lopez D. Ma W. Majoros J. McDaniel S. Murphy M. Newman T. Nguyen N. Nguyen M. Nodell S. Pan J. Peck M. Peterson W. Rowe R. Sanders J. Scott M. Simpson T. Smith A. Sprague T. Stockwell R. Turner E. Venter M. Wang M. Wen D. Wu M. Wu A. Xia A. Zandieh and X. Zhu. 2001. The sequence of the human genome. *Science* **291:** 1304-1351.

Yoder, J.A., C.P. Walsh, and T.H. Bestor. 1997. Cytosine methylation and the ecology of intragenomic parasites. *Trends Genet* **13:** 335-340.